Uncovering Mobile Infrastructure in Developing Countries with Crowdsourced Measurements

Uncovering Mobile Infrastructure in Developing Countries with Crowdsourced Measurements

Mah-Rukh Fida
The University of Edinburgh, UK
m.fida@sms.ed.ac.uk

Mahesh K. Marina
The University of Edinburgh, UK
mahesh@ed.ac.uk

ABSTRACT

Knowledge of cell tower locations enables multiple applications including identifying unserved or poorly served regions. We consider the problem of estimating the locations of cell towers using crowdsourced measurements, which is challenging due to the uncontrolled nature of the sample collection process. Using large-scale crowdsourced datasets from OpenCelliD with ground-truth cell tower locations, we find that none of the several commonly used localization algorithms (e.g., Weighted Centroid) nor the state of the art Filtered Weighted Centroid (FWC) approach that filters out less predictive measurements manage to deliver robust localization performance. We propose a novel supervised machine learning based approach termed as Adaptive Algorithm Selection (AAS) that adaptively selects the localization algorithm likely to provide the most accurate localization performance for a given cell and its crowdsourced samples. We show that AAS not only significantly outperforms the state-of-the-art FWC approach, with median error improvement over 65%, but also achieves localization performance within 20% of an idealized Oracle solution. We validate the applicability of AAS in new and different settings (including WLAN AP localization) before presenting case studies in three different African countries that demonstrate the use of AAS based cell tower localization to reliably infer mobile infrastructure in developing countries.

KEYWORDS

Cell tower localization, crowdsourced measurements

ACM Reference Format:
Mah-Rukh Fida and Mahesh K. Marina. 2019. Uncovering Mobile Infrastructure in Developing Countries with Crowdsourced Measurements. In *THE TENTH INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGIES AND DEVELOPMENT (ICTD '19), January 4–7, 2019, Ahmedabad, India.* ACM, New York, USA, 11 pages. https://doi.org/10.1145/3287098.3287113

1 INTRODUCTION

Understanding the deployment patterns of communication infrastructure in general offers several benefits, including improving competition and quality/cost of services in the telecommunication markets to the benefit of consumers. However network operators usually treat their infrastructure related information as sensitive from their market position standpoint and generally do not disclose it, except to regulators and policy makers and that too with a non-disclosure agreement. Same can be said about mobile communications network infrastructure, in particular the locations of cell towers. Even though the knowledge of cell tower locations allows external validation of operator provided mobile coverage maps and more crucially enable identification of unserved or poorly served regions by correlating with population data [19], it is rarely available in the public domain. So measuring from the outside (say, from user devices) and making inferences about the infrastructure is therefore the only means to estimate this information.

In view of the above, our focus in this paper is on estimating the cell tower locations from user-side measurements, especially crowdsourced samples. The above outlined uses clearly indicate the value of such estimation for developing country settings, which we highlight in this paper. However, generally speaking, knowing cell tower locations has several other use cases. Device localization via trilateration from multiple nearby cell towers is a popular application, offering an alternative to GPS when it is unavailable or for energy-efficiency reasons, as is evident from the cell tower location databases maintained by various location service providers [3, 5, 8]. Estimating cell footprint, reliably mapping coverage and finding/tracking density of cellular infrastructure are some other applications. There are benefits even for mobile network operators such as locating the transmitters from rogue networks operated using software-defined radio (SDR) platforms and getting insight on where to grow their infrastructure depending on the infrastructure owned by other operators.

While relying on measurements contributed by users is a cost-effective means for measurement-based cell tower localization, accuracy and robustness become challenging due to the lack of control over the measurement process on the device side. As elaborated in the next section, a number of algorithms are available in the literature for measurement based cell tower localization [2, 6, 7, 12, 13, 15, 20, 22, 23, 25], most of them emanated from the Wi-Fi access point (AP) localization context. Our analysis using a large-scale crowdsourced measurement dataset with ground-truth cell tower locations from OpenCelliD [3] reveals that none of these algorithms provide consistently good accuracy performance when used with crowdsourced measurements. Moreover we find that even the state-of-the-art cell tower localization approach, which we call as Filtered Weighted Centroid (FWC) [10], that filters out less predictive measurements from an accurate localization perspective is far from the best achievable localization outcome as it is limited by relying on a specific localization algorithm underneath.

Keeping in mind the above mentioned observations, we propose in this paper a novel approach called *Adaptive Algorithm Selection*

ICTD '19, January 4–7, 2019, Ahmedabad, India

ACM ISBN 978-1-4503-6122-4/19/01…$15.00
https://doi.org/10.1145/3287098.3287113

(AAS) to select the localization algorithm, from a suite of algorithms, that is expected to provide the most accurate cell tower localization for a given cell and an associated set of crowdsourced measurements. AAS employs a framework based on supervised machine learning for this purpose. Through an extensive measurement based evaluation, again using the OpenCelliD dataset, we show that our AAS approach significantly outperforms FWC [10] by more than 65% in median error and reduces the mean error by more than half. At the same time, AAS achieves median error under 20% of the Oracle scheme that always picks the best performing algorithm. In addition, we show that AAS provides similar improvements even when applied for Wi-Fi AP localization. Even more crucially, we examine the applicability of AAS model trained in one setting to other new settings for which there may not be ground-truth cell tower location information to retrain the model and obtain promising results.

In summary, this paper makes the following contributions:

- (Section 3) We examine the impact of crowdsourced measurement characteristics on cell tower localization using multiple datasets, and show that none of the commonly used algorithms nor the state of the art approach to filtering out less predictive measurements deliver robust localization performance.
- (Section 4) We propose a novel supervised machine learning based `Adaptive Algorithm Selection (AAS)` approach for robust cell tower localization with crowdsourced measurements. Using large-scale crowdsourced cellular and WLAN measurement datasets, we show that AAS not only significantly outperforms existing alternatives but also achieves localization performance within 20% of an idealized Oracle solution.
- (Section 5) Last but not least, we present three case studies in three different countries in Africa showing the use of AAS based cell tower localization to reliably infer mobile infrastructure in developing countries. This builds on the validation exercise we conduct examining the applicability of pre-trained AAS model for one setting in new settings (differing in operators and countries) where ground-truth cell tower location information is available for the latter.

The next section describes our datasets and metrics, and reviews the related work.

2 PRELIMINARIES

2.1 Datasets

We use four different types of datasets as listed in Table 1 with nearly 12 million measurement samples spanning over 15000 cells. Majority of our measurement data is from OpenCelliD [3], a community project aimed at building a database of cell towers around the world based on crowdsourced signal measurements. The reported cell tower locations are either ground-truth or aggregated from crowdsourced measurements. From this database we have used sub-datasets from Germany, Poland, Zambia, South Africa and Morocco. In most of these cases, each cell has at least 50 measurement samples; each sample corresponds to mobile network signal strength from a user device stamped with location, time, cell id (CID), radio access technology (RAT), etc. For the OpenCelliD datasets, we use measurements for various 2G variants (GSM, GPRS and EDGE) as they are the largest in number. Ground truth cell tower locations are only available for Germany and Poland. The other datasets are used for specific purposes towards motivation or demonstrating the generality of the proposed approach; as such, they are referred to in the respective sections of the paper.

2.2 Metrics

We mainly use two metrics to characterize the accuracy & robustness of cell tower localization approaches including our proposal: (1) *Mean absolute prediction error (MAPE)*, defined as the average Euclidean distance between estimated and ground truth locations of cell towers (or Wi-Fi APs), across all towers (APs); (2) *Median absolute prediction error*, defined similarly but focusing instead on the *median* of the errors. We also make use of box plots, bar charts and CDFs of localization errors in some cases to draw attention to the distribution of localization errors produced by different approaches.

2.3 Localization Algorithms

A number of algorithms have been used in the literature for estimation locations of cell towers from measurements, several of them have been originally proposed for Wi-Fi AP localization. Broadly speaking, these algorithms can be placed in four categories: angle-of-arrival, geometry, received signal strength (RSS) and propagation path loss based schemes. We consider a representative set of algorithms outlined in Table 2. Note that the choice of these particular set of algorithms is driven by their suitability of use with uncontrolled crowdsourced measurements collected by commodity devices [10, 14]. For this general reason, we do not consider angle-of-arrival/directionality based approaches such as DrivebyLoc [20] and Borealis [25]; coarse-grained approaches that give zip code level estimates for transmitter location such as in [15]; as well as approaches that require meticulous orchestration of measurement collection as in CrowdWiFi [22]. Among the seven algorithms in Table 2, we find that MEC and GPR generally result in relatively high localization errors as shown in Fig. 1. So in the rest of the paper, we limit our focus to the remaining five algorithms along with recent work in [10] and our proposed approach.

3 MOTIVATION

3.1 Impact of Crowdsourced Measurement Characteristics

Crowdsourced measurements are uncontrolled as they are reported from random locations at diverse environmental situations, times and devices. Here we state some of the characteristics of crowdsourced measurements that either favor or hurt cell tower localization accuracy of a given algorithm, motivating the need for our proposed adaptive algorithm selection approach.

Inaccuracy in measurement location. We use the RF Signal Tracker dataset to assess the impact of measurement location inaccuracy on a transmitter's predicted position. This dataset consists two sets of measurements for same route but one collected with GPS and other with network-based location information. The inaccuracy for GPS-based measurement locations ranges from 8m-16m with median value of 12m while for network based locations it

Table 1: Summary of datasets.

Dataset	Description	# Cells	#Samples
OpenCelliD	MCC 262 (Germany), MNC 01 (T-mobile)	2002	1,579,120
	MCC 262 (Germany), MNC 02 (Vodafone D2)	4200	3,378,117
	MCC 260 (Poland), MNC 01 (Polkomtel)	2000	1,164,741
	MCC 645 (Zambia), MNC 01 (Airtel)	2000	2,453,827
	MCC 655 (South Africa) MNC 07(Cell C)	5379	2,766,26
	MCC 604 (Morocco) MNC 01 (IAM)	1676	80,213
RF Signal Tracker dataset	Collected with Samsung Galaxy S3 during 18-20 April 2017 in Edinburgh city center.	68	6,000
Synthetic dataset	Generated using Okumara-Hata model [11] with range 35Km, carrier frequency 1700 MHz, antenna of height 200m and end-device at 3m depicting a cell in small and medium-size cities.	1	maximum 32,000
Dartmouth WLAN dataset [9]	Collected at Dartmouth campus using Place Lab software during 12–14 Sept 2005. We used warwalk dataset with at least 20 samples per AP.	280 APs	31,312

Table 2: Existing cell tower localization algorithms suitable for crowdsourced measurements.

Algorithm	Brief Description
Geometric schemes	
Centroid (C) [6, 12]	Mean of all measurement locations.
Weighted Centroid (WC) [6, 12, 23]	Weighted average of measurement locations using signal strengths at each location as a weight.
Minimum Enclosing Circle (MEC) [12]	Center of the circle with minimum radius that encloses all measurements.
Received Signal Strength (RSS) based schemes	
Strongest RSS (SRSS) [23]	Picks the location with the strongest RSS measurement sample for a cell.
Grid Likelihood (GL) [12, 13]	It imposes a grid structure on the region covered by a base station and picks the center of the grid cell with the highest likelihood of receiving maximum RSS as the estimated cell tower location.
Gaussian Process Regression (GPR) [2]	Here the signal propagation from the base station is modeled via Gaussian process regression with the peak signal strength location from the model used as the estimated cell tower location.
Propagation Path Loss (PPL) based schemes	
Monte Carlo Path Loss (MCPL) [7]	It uses a grid structure, like in [2], with path loss coefficient estimated for each grid cell using the available signal strength measurements. The (center of) grid cell that results in the least squared error between estimated and actual signal strength measurements is estimated as the cell tower location.

ranges from 20m-40m with median of 22m. With these two sub-datasets, we observe in Table 3 that inaccuracy in measurement locations adversely affects all types of localization algorithms but

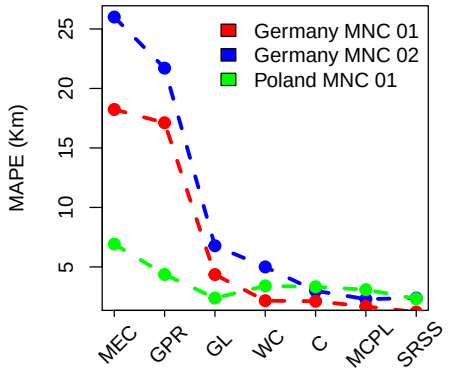


Figure 1: Error performance of different localization algorithms using OpenCelliD datasets with ground truth cell tower locations (Table 1).

to different degrees. RSS and PPL based schemes seem to be much more impacted compared to geometric algorithms. This is because geometric schemes take into account overall spread of measurements and are not sensitive to location errors unless center changes. Localization accuracy with MCPL, on the other hand, degrades the most because of non-alignment of samples' distances from probable cell tower location and their path loss values.

Layout of measurement samples. Crowdsourced measurements come in different layouts. For example, samples generated by pedestrians and passengers are along streets and roads while those at hotspots, homes and work places may have more random locations. These layouts are highly dependent upon deployment location of a cell, surrounding landscape, crowd population density, their moving patterns in the cell, and a cell's footprint.

Fig. 2 shows some of these layouts generated with the synthetic dataset mentioned in Table 1. The *well spread* case shown in Fig. 2 (a) is a good layout for each of the localization categories and it also exhibits high negative correlation between samples' distances to cell site and their corresponding RSSs. When the layout is *skewed* as shown in Fig. 2 (b), localization errors for schemes in the geometric

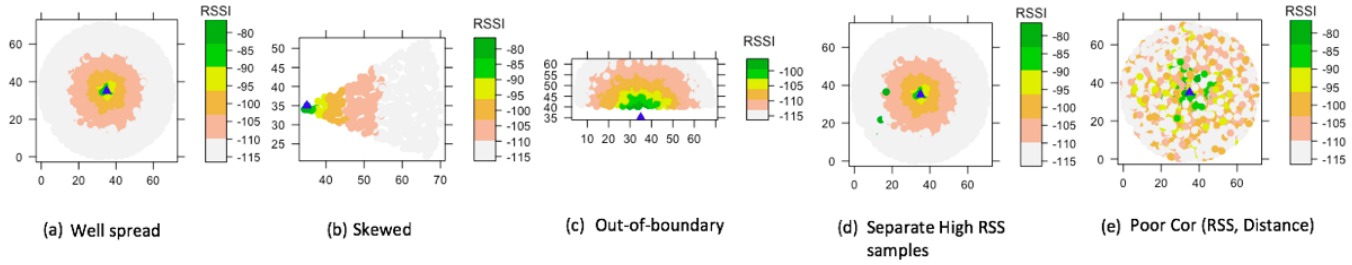


Figure 2: Different spatial layout of measurements.

Table 3: Effect of crowdsourced measurement characteristics on prediction error (in meters).

Characteristic	Geometric		RSS		PPL
Schemes	C	WC	SRSS	GL	MCPL
Location inaccuracy impact with RF Signal Tracker dataset					
GPS locations	261	263	238	238	227
Network-based	265	266	264	277	291
% degradation	1.5	1.1	10.8	16.3	28
Impact of measurement layouts with Synthetic dataset					
Well spread	352	356	754	707	388
Skewed	22K	22K	754	1K	1K
Out-of-boundary cell	16K	16K	5K	5K	5K
High RSS samples separated	352	360	18K	707	668
Impact due to Cor(RSS, distance to cell tower) with Synthetic dataset					
Poor correlation	352	434	2K	2K	1K
% degradation	0	21	144	182	222

category are high (Table 3) as they incline towards the center of the sampled region. Directional tower up on a hill, restricted building or like features make measurements to be reported away from the cell tower location leading to *out of the boundary* (Fig. 2 (c)) layout. Such a layout is poor for all localization algorithms though to different degrees as each algorithm estimates cell location to be lying somewhere inside the measurement boundary. Being a worse layout for all localization approaches J. Yang et. al [23] introduced "Boundary Filtering" technique that filters out such measurement scenarios to be not included in cell localization process. A favorable layout for RSS based approaches is when there is a single RSS peak while the opposite is true when there are multiple separated peaks shown in Fig. 2 (d). Such a *Separate High RSS samples* layout makes the tower location prediction erroneous especially for SRSS as indicated by results in Table 3.

Correlation of RSS to distance from cell tower. With free-space path-loss, signal strength is inversely proportional to the square of the distance from the transmitter. This ideal relation does not always hold because of effects (reflection, etc.) from the ground and objects in the path. Algorithms highly dependent upon samples' signal strengths are impacted adversely as the correlation between samples' signal strength and distance to cell tower weakens. Using the synthetic dataset visualized in Fig. 2 (e) we observe from results in Table 3 that, when compared to *well spread* layout, the percentage degradation in accuracy is substantial for the algorithms relying on signal quality of the sampled locations. In contrast, accuracy with the Centroid algorithm remains unaffected. WC also observes a dip in accuracy as it takes RSS as a weight for cell tower locations.

Other than above features, the localization algorithms are impacted by different degrees with the outliers in measurement locations due to cell dragging and cell-on wheels [16] and un-reliability in signal values due to diverse user devices. Note that we do not explicitly handle the issue of diversity among user devices contributing measurements and instead refer the reader to our other paper [4] that specifically studies the impact of device diversity.

Is there a clear winner among the commonly used measurement based cell tower localization algorithms? We now examine the overall error performance of five commonly used cell tower / AP localization algorithms (i.e., the ones in Table 2 except MEC and GPR) over three real-world crowdsourced measurement datasets: OpenCelliD datasets for Germany MNC01 and Poland MNC01; and Dartmouth WLAN dataset (Table 1). For this analysis and henceforth, *a measurement scenario* is a set of measurements (equivalently, samples) available for a cell tower (AP). For each measurement scenario in each of the three above mentioned datasets, we apply each of the five algorithms and calculate the percentage of measurement scenarios in a dataset when each algorithm provides the least localization error. If there was a clear winner among the algorithms, then that algorithm would have a percentage of 100 at the expense of other algorithms. Fig. 3 shows the results in the form of pie charts. While SRSS is the best performing one for majority of measurement scenarios in all three datasets, it performs really poorly when it is not the best. Specifically, SRSS has a long tail with some extremely high errors (for example as Fig. 4 (c) displays an error above 10Km) that lead to higher mean and standard deviation of errors (above 1.5Km and 5.5Km, respectively). In summary, we conclude that there is no single consistently best performing algorithm from among the commonly used localization algorithms.

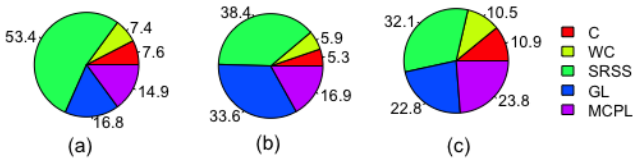


Figure 3: Percentage of measurement scenarios in (a) OpenCelliD Germany MNC 01; (b) OpenCelliD Poland MNC 01; and (c) Darmouth WLAN datasets for which each of the five commonly used localization algorithms performs best.

3.2 Limitations of Existing Localization Approaches to Deal with Crowdsourced Measurement Characteristics

As the peculiar characteristics, noise and outliers in crowdsourced measurements can negatively impact the accuracy of any given wireless infrastructure localization algorithm, some research studies employ pre-processing of measurements before application of a localization algorithm. For example to improve localization accuracy of SRSS and WC, J. Yang et. al [24] introduced three pre-processing steps, namely *RSS Thresholding*, *Boundary Filtering*, and *Tower-based Regrouping*. Concerning rationale behind the first step, the authors in [24] argue that once the strongest observed RSS drops below -60 dBm, the localization error of the SRSS algorithm increases significantly due to drop in correlation between the strongest RSS and the distance to the cell tower. Secondly if such a sample lies at the boundary of measurement layout, they suggest to exclude such samples. However we observe in our datasets that for 50%of the cases where maximum RSS is below -60 dBm, SRSS localizes cells within 500m error, which is not a very high error for cell tower localization. Moreover, rather than exclusion we believe that coarse estimation is better than no estimation as it provides a rough idea of probable deployment area of a network's infrastructure. For the third step, authors in [24] claim that merging measurements of cell sectors with same cell tower improves WC results as it removes the ill-effect of skewed measurements. Merging cell sectors is, however, beneficial only when one knows naming pattern of CID's associated to sectors of same cell tower. In an other study, to have a reliable analysis out of crowdsourced data, F. Ricciato et al. [16] presented a few solutions to some crowdsourced measurement issues including identification of erroneous cell-IDs, unrealistic cell sizes, effect of antenna dragging and outliers.

To get the most out of a crowdsourced measurement dataset, the work of Zhijing Li et al. [10] can be regarded as the most recent work. It assesses the predictive value of a subset of measurement samples and finds that samples with high RSS standard deviation (> 100k) and low RSS-weighted dispersion mean (< 0.5km) correlate to high localization accuracy for WC. Based on this observation, they devised a variant of WC which we call as `Filtered WC (FWC)`; it relies on measurements that meet either or both of the above two criteria. To see if FWC offers a satisfactory alternative to the five algorithms studied in Fig. 3 above, Fig. 4 (a) shows a measurement scenario where FWC chooses a smaller subset of samples as predictive with RSS-weighted dispersion mean of 635m and standard deviation of RSS double to that of the whole measurement set. While filtering measurements can be useful sometimes, there are also pitfalls underlying the FWC approach:

- One has to iteratively collect more measurement samples until RSS-weighted dispersion mean threshold and high standard deviation of RSS samples are met, which may not be practical if given a dataset that does not meet both of these criteria.
- If a measurement subset meets either of the two criteria used in FWC, it is not always true that the left out measurement samples perform poorer with traditional WC approach.
- Finally FWC bases its localization on a single algorithm (i.e. WC). As we saw in Fig. 3, none of the algorithms is clearly superior over others. This can be further verified with localization errors in Fig. 4 (b), where another algorithm, SRSS, exploits the available measurement samples better than the FWC algorithm.

4 TOWARDS ROBUST CELL TOWER LOCALIZATION

Results and discussion from the previous section show that relying on a particular localization algorithm or filtering out measurement samples both limit localization accuracy that can be achieved for a given measurement scenario. So in this section we propose an alternative novel paradigm which is to use all available measurements and choosing between different localization algorithms (for example, the ones studied in Fig. 3). This paradigm is in sharp contrast to the approach taken in FWC [10] where the algorithm to be used is fixed first (WC) and then the measurements are filtered to retain only those that are likely to help in achieving high localization accuracy. Note that we do not filter out any samples with the rationale that the algorithm if carefully chosen can exploit all the available measurement samples.

4.1 Probabilistic Algorithm Selection

We first examine a naive instantiation of this new paradigm to choose between different algorithms. Based on available measurement data for a set of cells (measurement scenarios) and corresponding ground-truth cell tower location information that can be viewed as *training data*, we estimate the relative percentage of scenarios in which each algorithm yields the best localization result as in Fig. 3. Such a pie chart is used for all subsequent cell tower location estimations for probabilistically choosing an algorithm; essentially, percentages (divided by 100) in the pie chart serve as prior probabilities for picking different algorithms. We refer to this approach as `Probabilistic Algorithm Selection (PAS)`.

Using 10-fold cross-validation [17] on the three OpenCelliD datasets from Germany and Poland, we see that PAS outperforms FWC by 46 to 65% in the different datasets as shown in Fig. 7. It is however off from the "Oracle" results by more than 57% both in mean and median errors. By Oracle results, we refer to the best estimated cell tower locations, among that estimated by the five individual algorithms, for each cell (measurement scenario). The results in Fig. 7 also show that PAS performs poorly compared to SRSS, the algorithm that yields the lowest localization error in majority of the scenarios. Moreover, median Oracle results for the three datasets are 422 m, 300 m and 558 m better than that of

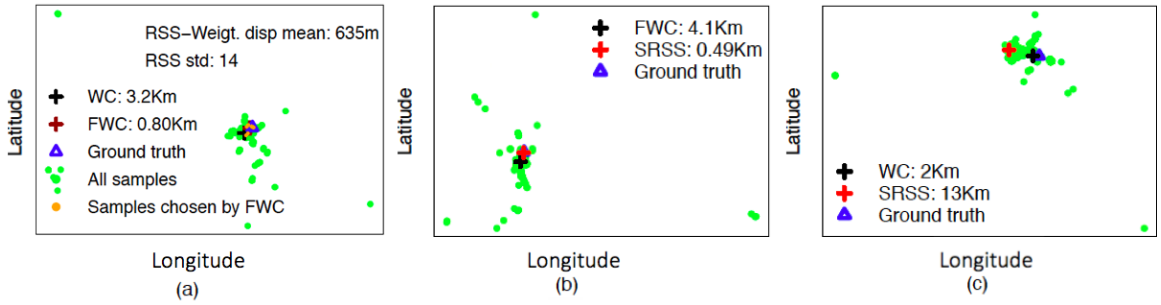


Figure 4: Example measurement scenarios that show (a) FWC outperforming WC with fewer carefully chosen samples; (b) SRSS yielding a significantly better localization accuracy than FWC, and (c) SRSS producing a very high error.

PAS by a large margin, indicating that localization accuracy can significantly improve if correct localization algorithm is chosen.

4.2 Adaptive Algorithm Selection

The results from the previous subsection suggest that while the simple-minded PAS (reflecting the approach to choose between algorithms) is already better than FWC (that is based on a specific algorithm – WC) it leaves room for substantial improvement compared to Oracle and SRSS.

In light of the above, we propose a more sophisticated variant called `Adaptive Algorithm Selection (AAS)`. AAS views the problem of choosing a localization algorithm from among the suite of different algorithms as a classification problem in machine learning – different algorithms make up different classes for the classifier. Unlike PAS which somewhat randomly selects an algorithm with no regard to the specific characteristics of the measurement scenario for which cell tower needs to be localized, AAS classifier model considers a variety of features (outlined next) that aid in distinguishing between different measurement scenarios and algorithms.

4.2.1 Feature Set. In Table 3, we illustrate some of the features with varying degree of impact upon the localization algorithms. For an in-depth understanding of the combination of features that can serve as a guide for assessing the suitability of a localization algorithm for a given measurement scenario, we extract four types of features as listed below:

- **Measurement Spread Features:** These include size (number of measurement samples); radius (spatial spread of the samples as determined by the radius of the minimum enclosing circle); DistTl (mean distance of all samples to the "trend" line of the samples); DispCenter statistics (i.e., mean, median, standard deviation and index of dispersion of the samples from central location of minimum enclosing circle); and Density (mean number of samples per sq. km for a cell).
- **Signal Strength Features:** These consist of RSS statistics as well as highest RSS statistics. For the latter, we use number of highest RSS samples; minimum, maximum, mean and standard deviation of distances among highest RSS sample locations.
- **Weighted Measurement Spread Features:** These include DispRSS statistics (i.e. mean, median, standard deviation and index of RSS-weighted dispersion from central location) and Autocorrelation among samples.
- **Features based on Estimated Locations:** These extract correlation between signal quality of measurements and their distances to estimated locations of the five algorithms in Fig. 3; distance between each pair of estimated locations and distance of each estimated location to the center of the trend line.

4.2.2 AAS Model Generation. We take a supervised machine learning approach to the classification problem stated above. To generate the AAS classifier model, for a subset of measurement scenarios with ground-truth cell tower location information, we train the model as follows. For each of these measurement scenarios (cells), we create tuples with features computed as in the previous subsection and the algorithm among the five in Fig. 3 that yields the least localization error as the class label.

Another key question to realize the classification model is the selection of a classification technique that yields the most accurate classification. We empirically address this question and compare the accuracy with seven well-known and commonly used classification techniques: K Nearest Neighbors (KNN), Naive Bayes (NB), Decision Tree (DT), Multinomial Regression (MNReg), Neural Networks (NNET), Support Vector Machine (SVM) and Random Forest (RF). Results shown in Fig. 5 indicate RF to be the best technique with accuracy ranging between 56 and 73% for different datasets, so we use RF as the classification technique in AAS. From deeper examination, we find that there are two main reasons for the somewhat low level of classifier accuracy: (1) imbalance between the different classes, indicated earlier by Fig. 3; and (2) confusion between subsets of algorithms (classes) having similar localization inaccuracy within few tens of meters of each other. Even so, as we will see later in this section, the AAS performs significantly better than the state of the art and the simple-minded PAS.

Significant Features. As for the significant features, AAS model is highly impacted by features showing mutual distance gap between the estimated cell tower locations of different localization algorithms. Table 4 shows the importance of the top impacting features in the form of Mean Decrease in Accuracy (MDA). The more the accuracy of an RF-based classifier decreases due to the exclusion (or permutation) of a particular feature, more important

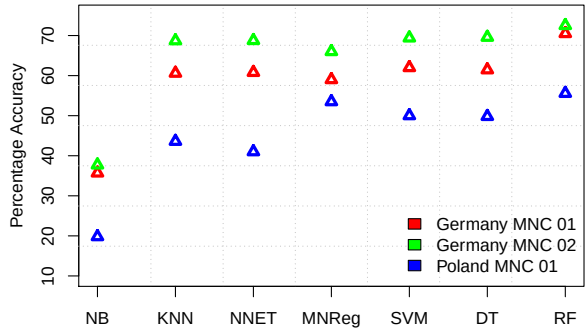


Figure 5: Comparison of accuracy between various classification techniques that can be used to generate AAS model.

that feature is deemed, and therefore features with a large MDA are more important for the purpose of classification.

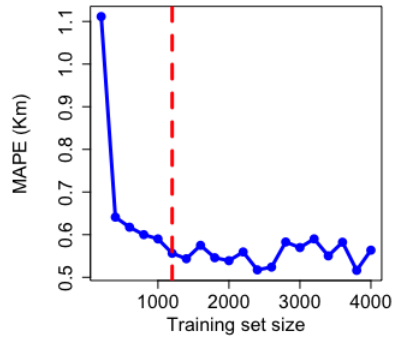


Figure 6: A training set size of around 1200 is sufficient for AAS to deliver low localization error.

4.2.3 AAS Evaluation. We now evaluate the localization performance obtained with AAS in comparison with Oracle and the other alternatives discussed before. Our evaluations are based on two methods: (1) 10-fold cross validation (CV); and (2) using a training set of around 1200 scenarios and test set of 200 scenarios. The selection of training and test sets is random with results an average of ten runs. We choose training set size to be 1200 scenarios as the learning curve given in Fig. 6 suggests this training set size is sufficient to train an AAS model.

Due to space restrictions, we only present the results from 10-fold CV in Fig. 7 (a). Because of its more reliable choice of a localization algorithm, AAS reduces the median localization error by 42.4%, 28% and 25.7% respectively for the three datasets, compared to PAS. For the same reason, the median localization accuracy with AAS is within 20% of the Oracle performance in all three crowdsourced datasets.

4.2.4 AAS Applicability to WLAN AP Localization. Given that many of the algorithms employed for cell tower localization have originally been proposed for localizing the Wi-Fi APs (e.g., [2, 6, 7]), it is natural to wonder if an approach like AAS that is seen to be effective for cell tower localization is also effective for the AP localization setting. To address this question, we use the Dartmouth WLAN dataset and compare the different schemes.

Table 4: Features for AAS model in decreasing order of importance with RF as the classification technique and as per MDA.

Features	MDA
Distance between estimated locations of the algorithms	30
Maximum RSS value	24
Distance between estimated locations of algorithms and trend line	23.6
Mean RSS value	23.3
Autocorrelation of samples' locations and their RSS values	22
Mean dispersion of samples from central location	21.7
RSS standard deviation	20.9
Standard deviation of RSS weighted dispersion of samples	20.6
Mean distance of samples to trend line	20.2
Size of measurement samples	20
Median of RSS weighted dispersion of samples from central location	19
Correlation of samples' distances and RSS values to Geometric locations	18.5
Median of dispersion of samples from central location	18

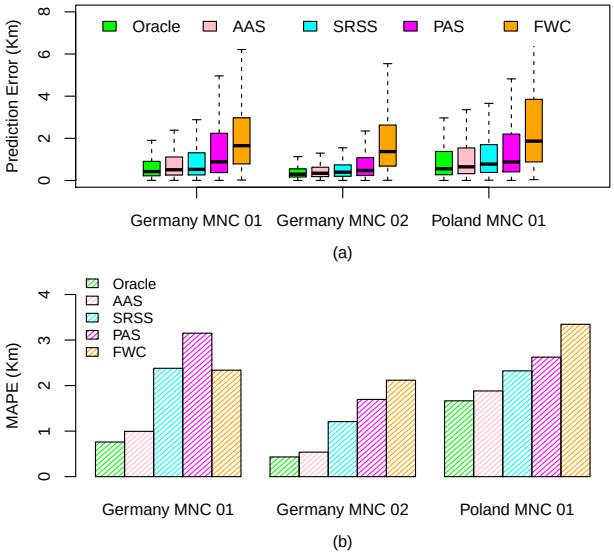


Figure 7: (a) Distribution of localization errors and (b) mean absolute prediction error (MAPE) with different schemes using OpenCelliD datasets.

Results in Fig. 8 (a) indicate similar relative performance as before. However, different from the cell tower localization setting, the ratio of improvement in this setting is lower for AAS. We find that the dataset used is the key reason behind these observations. Dartmouth dataset is not composed of crowdsourced measurements; instead it is collected via war-driving and war-walking restricting to roads. Moreover, this dataset is relatively smaller in size compared to

previously used OpenCelliD dataset, both in terms of the number of measurements and scenarios (280 APs vs. 2000-4200 cell sites). Nevertheless, these results do clearly demonstrate the robustness of the AAS approach across different wireless infrastructure localization settings.

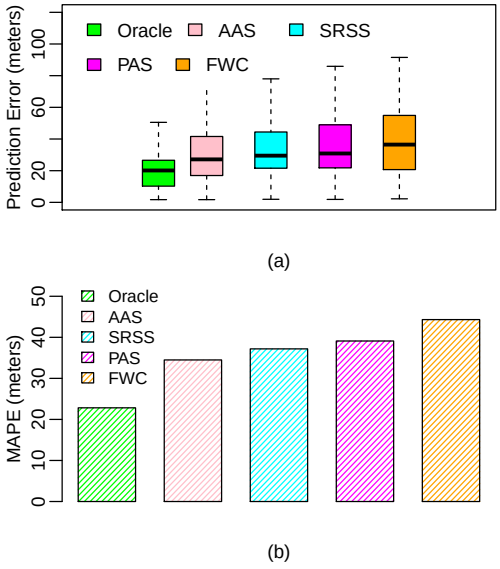


Figure 8: (a) Distribution of localization errors and (b) MAPE with AAS in comparison to other schemes using WLAN dataset.

5 AAS APPLICABILITY IN NEW AND DIVERSE SETTINGS

As stated at the outset, a key motivation behind our study into localizing cell towers with measurements is to have a means to gain insight into the reach of mobile infrastructure in developing country settings. To this end, we initially validate AAS model in new settings before demonstrating its applicability in developing country settings under the typical and realistic assumption that in such cases ground-truth cell tower location information is unavailable.

5.1 AAS Validation in New Settings with Ground Truth

Here we first examine if AAS model trained with data from a particular operator and country can be used in a different setting (different country and operator) assuming ground truth cell tower location information is available for the latter.

This investigation is aimed at testing the applicability of the AAS model in new and previously unseen settings. For this purpose, we use Germany's MNC 01 dataset to train our AAS model and test it over Germany's MNC 02 and Poland's MNC 01 datasets; we refer to this variant of AAS as `AAS (Diff.)`. For comparison, we also include AAS variant which is trained and tested on different parts of the same dataset (e.g., Germany's MNC 02) as well as Oracle and FWC.

Table 5 and Fig. 9 show the results. The focus is on the difference in error performance between the two AAS variants, `AAS (Same)` and `AAS (Diff.)`, former indicating the best case result achievable with AAS in a new setting. Results indicate that the difference between these variants is marginal with both test datasets and close to the Oracle performance, and that `AAS (Diff.)` is significantly better than FWC or SRSS.

Moreover one should look at the distribution of features' values that provide guidance about the extent an AAS model trained on different dataset is trustable. For example for the case of Poland MNC 01, we observe its measurement scenarios' radii, mean dispersion of samples from center, autocorrelation and distances of algorithms' results (from each other, to trend line, and to center) are comparatively smaller while minimum and median RSS values are higher. All of these attributes are significant features effecting decision of AAS model where some of these distributions are shown in Fig. 10. In other words, smaller the difference in distribution of features' values, from training dataset, higher is the overall localization accuracy achieved by `AAS (Diff.)`. Other than visual inspections one can use Chi-square homogeneity [21] type of tests to investigate if difference between two set of distributions is non-significant.

Due to difference of the distribution we initially obtained a difference of 6% in the median error between `AAS (Diff.)` and `AAS (Same)` for the Poland's dataset. By dropping some of the drifting variables [18] that cause highest covariate shift and using instances from training set similar in distribution to that of test set, the error dropped by 2%. Features with covariate-shift are the ones having very different distribution for both the test and the train sets. While dropping these features care should be taken not to remove the highly significant ones. Secondly while generating the model, accuracy can be improved by either assigning higher weight or retaining the instances from training set that are similar to those in the test dataset.

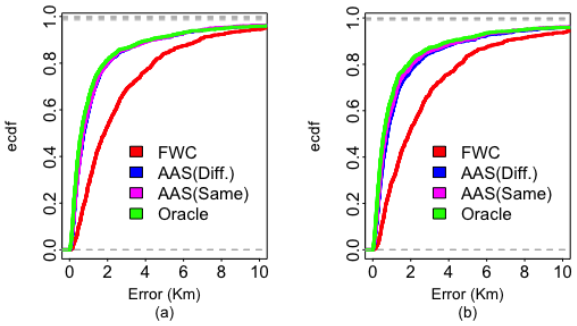


Figure 9: Localization errors with `AAS (Diff.)` that is trained on Germany MNC 01 dataset and tested on (a) Germany MNC 02 and (b) Poland MNC 01 datasets, relative to other schemes.

5.2 AAS Application to Developing Country Settings

Now we come to our key motivating use case of estimating cell tower locations in new settings where there is no ground-truth

Table 5: Key localization error statistics with `AAS (Diff.)` that is trained on one dataset (Germany MNC 01) and tested on different datasets, compared to others schemes and with respect to Oracle.

Test Set	Germany MNC 02		Poland MNC 01	
Scheme	*Median APE*	*Mean APE*	*Median APE*	*Mean APE*
Oracle	309m	431m	547m	1.62km
AAS(Same)	11.6%	18.5%	18%	17%
AAS(Diff.)	12.6%	27%	22%	19%
SRSS	14%	148%	32%	38%
FWC	335%	386%	241%	136%

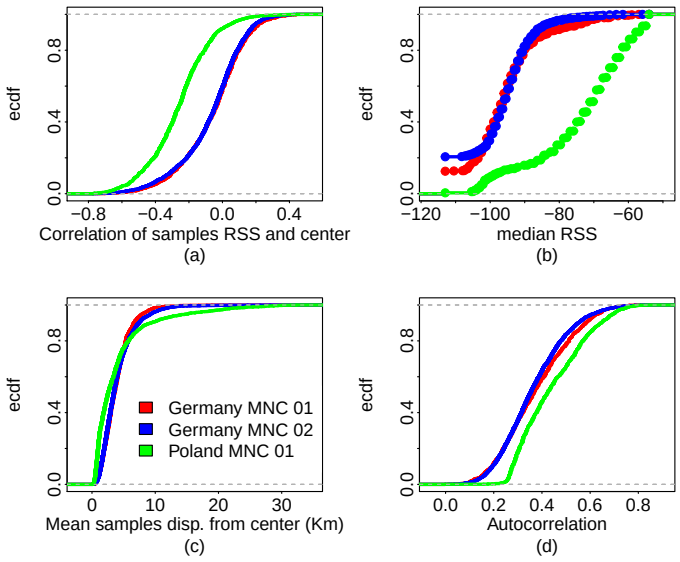


Figure 10: Differences in the distributions of some features used by AAS across different datasets that can impact accuracy of `AAS (Diff.)`.

information available, keeping developing countries in mind. Results from the preceding subsection suggest that a pre-trained AAS model when used in an entirely different setting still gives location estimations within around 20% of the Oracle approach, which makes it reasonably trustworthy and that too by a big margin compared to the alternative schemes from the literature. We can do even better by taking into account the specific characteristics of measurement features for the target setting and accordingly choosing a model trained for a similar setting. As demonstrated in Fig. 10, different settings differ in their measurement characteristics. We exploit this observation in our case studies below.

To demonstrate the usefulness of AAS in inferring mobile infrastructure in developing countries via measurements, we consider three representative countries from Africa – Zambia, South Africa and Morocco – as case studies. We selected these countries keeping in mind availability of crowdsourced measurements in the OpenCelliD dataset and side information in the form of some publicly available coverage maps to visually inspect and assess the correctness of cell tower location estimations made by AAS.

Considering our first case study of Zambia, we focus on cell tower infrastructure for Airtel (MNC 01), which is one of the three largest operators in the country but does not even provide coverage map on its official website [1] let alone revealing its infrastructure siting information. For crowdsourced measurements for this operator, we rely on OpenCelliD's sub-dataset for Zambia. To apply AAS in this new setting, we train it on Poland MNC 01 dataset (in view of its similarity in distribution of features to that seen in Zambia). Resulting cell tower location estimations indicate the probable infrastructure layout of this operator (Fig. 11 (a)), which shows good alignment with the crowdsourced measurement based coverage map information available for this operator from OpenSignal (Fig. 11 (b)).

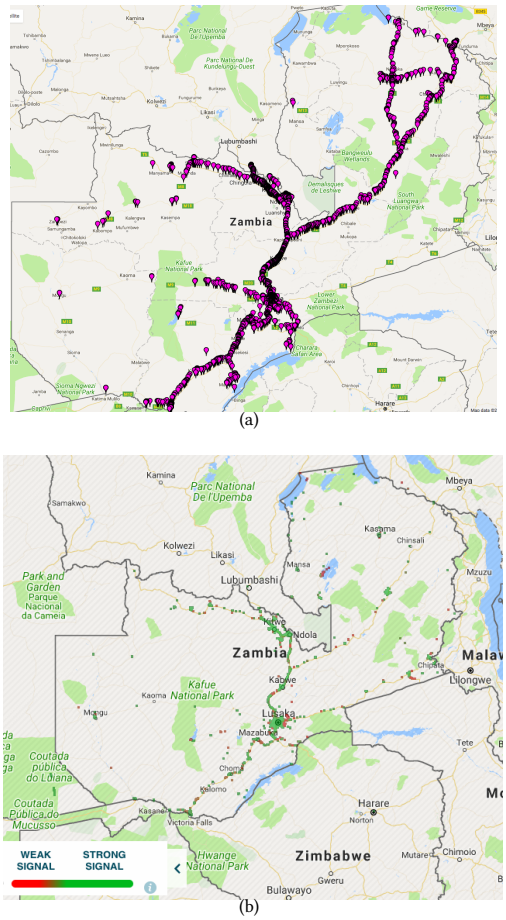
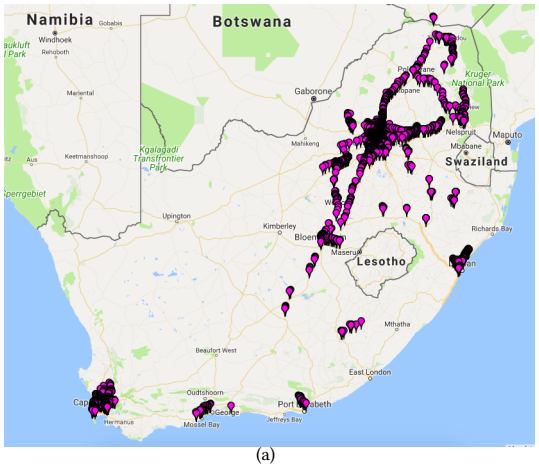


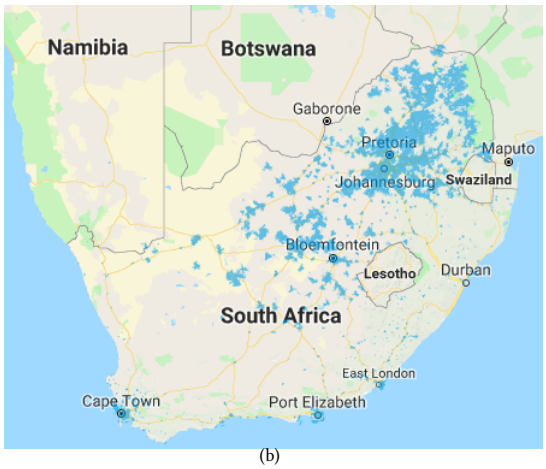
Figure 11: (a) `AAS` model trained on OpenCelliD Poland MNC 01 dataset and tested over measurement data for Airtel MNC 01 in Zambia from OpenCelliD; (b) Publicly available coverage status for Airtel, Zambia from OpenSignal.

We repeated a similar process of estimating cell tower locations for CellC (MNC 07) 2G mobile network in South Africa and IAM (MNC 01) network in Morocco. For both cases, we trained the AAS

model on Germany MNC 01 dataset in view of its feature similarity to the above test networks, like above. Resulting map with inferred cell tower locations for both these networks in different countries along with corresponding but independent coverage maps from public sources adding confidence to these inferences are shown in Figs. 12 and 13, respectively. These case studies clearly demonstrate the value of AAS approach for robust measurement based cell tower localization to map/track mobile infrastructure in developing countries.

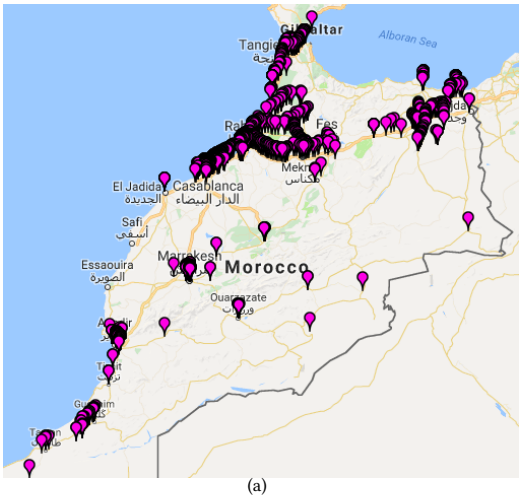


(a)

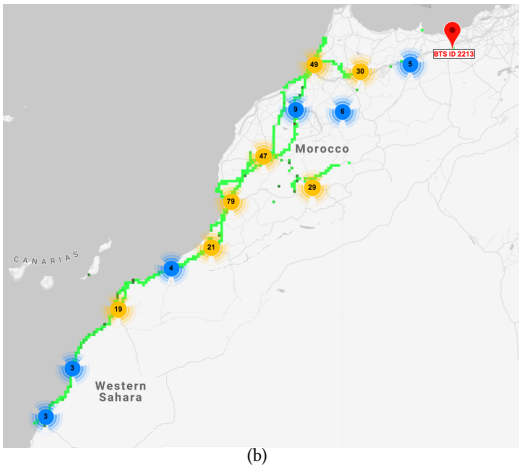


(b)

Figure 12: (a) Infrastructure layout of South Africa's CellC 2G network as identified by AAS (Diff.) with measurement samples obtained from OpenCelliD and (b) Coverage map of South Africa's CellC 2G network from its official website.



(a)



(b)

Figure 13: (a) AAS (Diff.) identifies infrastructure layout of GSM cells of Morocco's IAM network using measurement dataset from OpenCelliD and (b) Infrastructure layout of the same network as shown by cellmapper.com.

6 CONCLUSIONS

In this paper, we have considered cell tower localization using crowdsourced signal strength measurements. Using a large-scale crowdsourced dataset with ground-truth cell tower locations, we first showed that each of the commonly used localization algorithms is susceptible given the wide variations in features across different measurement scenarios. Even the recent FWC approach [10] to avoid using less predictive measurements in conjunction with a specific algorithm is found to be similarly vulnerable, making it produce high localization errors. Motivated by these observations, we proposed AAS, a novel localization approach based on supervised machine learning, aiming to adaptively select the localization algorithm that is expected to yield a most accurate localization performance for a given measurement scenario. AAS not only significantly outperforms the commonly used algorithms and the FWC approach but is also shown to be robust across new and different settings including WLAN AP localization. More crucially, we present case studies highlighting the use of AAS based cell tower localization in three different African countries to demonstrates its use for infering mobile infrastructure in developing countries.

REFERENCES

[1] Airtel. [n. d.]. Airtel. http://www.africa.airtel.com/wps/wcm/connect/africarevamp/Zambia. ([n. d.]). visited June 2018.

[2] Youngsu Cho et al. 2012. Improved Wi-Fi AP position estimation using regression based approach. In *Proc. Int'l Conf. on Indoor Positioning and Indoor Navigation (IPIN).*

[3] ENAiKOON. [n. d.]. OpenCelliD. http://www.opencellid.org. ([n. d.]). visited July 2016.

[4] M. Fida and M. K. Marina. 2018. Impact of Device Diversity on Crowdsourced Mobile Coverage Maps. In *Proc. 14th Int'l Conf. on Network and Service Management (CNSM'18).*

[5] Glenn Fleishman. [n. d.]. How the iPhone knows where you are. http://www.macworld.com/article/1159528/smartphones/how-iphone-location-works.html. ([n. d.]). visited July 2017.

[6] Dongsu Han et al. 2009. Access Point Localization Using Local Signal Strength Gradient. In *Proc. Passive and Active Network Measurement (PAM) Conference.*

[7] Myungin Ji et al. 2013. A Novel Wi-Fi AP Localization Method Using Monte Carlo Path-loss Model Fitting Simulation. In *Proc. IEEE PIMRC.*

[8] Zhengrong Ji and Ravi Jain. [n. d.]. Google enables Location-aware Applications for 3rd Party Developers. http://googlemobile.blogspot.co.uk/2008/06/google-enables-location-aware.html. ([n. d.]). visited June 2017.

[9] Minkyong Kim, David Kotz, and Jeffrey J. Fielding. 2006. CRAWDAD dataset dartmouth/wardriving. http://crawdad.org/dartmouth/wardriving/20060602. (2006).

[10] Zhijing Li, Ana Nika, Xinyi Zhang, Yanzi Zhu, Yuanshun Yao, Ben Y. Zhao, and Haitao Zheng. 2017. Identifying Value in Crowdsourced Wireless Signal Measurements. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017.* 607–616.

[11] Andreas F. Molisch. 2011. The Okumura-Hata Model. https://www.wiley.com/legacy/wileychi/molisch/supp2/appendices/c07_Appendices.pdf. (2011).

[12] Eric Neidhardt, Abdulbaki Uzun, Ulrich Bareth, and Axel Kupper. 2013. Estimating Locations and Coverage Areas of Mobile Network Cells based on Crowdsourced Data. In *Proc. 6th Joint IFIP Wireless and Mobile Networking Conference (WMNC'13).*

[13] Petteri Nurmi, Sourav Bhattacharya, and Joonas Kukkonen. 2010. A Grid-Based Algorithm for On-Device GSM Positioning. In *Proc. ACM UbiComp.*

[14] Henri Nurminen, Marzieh Dashti, and Robert Piché. 2017. A Survey on Wireless Transmitter Localization Using Signal Strength Measurements. *Wireless Communications and Mobile Computing* 2017 (2017).

[15] Gyan Ranjan et al. 2011. Unzipping Cellular Infrastructure Locations via User Geo-Intent. In *Proc. IEEE INFOCOM.*

[16] F. Ricciato, P. Widhalm, M. Craglia, and F. Pantisano. 2015. Estimating Population Density Distribution from Network-based Mobile Phone Data. *JRC Technical Report* (2015).

[17] Jeff Schneider. 1997. Cross Validation. https://www.cs.cmu.edu/~schneide/tut5/node42.html. (1997).

[18] Shikhar. [n. d.]. Train Test Similarity. https://www.kaggle.com/shikhar1/train-test-similarity. ([n. d.]). visited July 2017.

[19] Steve Song. 2017. Mapping the Unserved. https://manypossibilities.net/2017/04/mapping-the-unserved/. (April 2017). visited July 2018.

[20] Anand Prabhu Subramanian, Pralhad Deshpande, Jie Gao, and Samir R. Das. 2008. Drive-By Localization of Roadside WiFi Networks. In *Proc IEEE INFOCOM.*

[21] Anthony Tanbakuchi. 19 May 2009. Tests of independence and homogeneity. http://www.u.arizona.edu/~kuchi/Courses/MAT167/Files/LH_LEC.0640.HypTest.IndepHomog.pdf. (19 May 2009).

[22] Di Wu et al. 2014. CrowdWiFi: Efficient Crowdsensing of Roadside WiFi Networks. In *Proc. ACM Middleware Conference.*

[23] Jie Yang et al. 2010. Accuracy Characterization of Cell Tower Localization. In *Proc. ACM UbiComp.*

[24] Jie Yang, Alexander Varshavsky, Hongbo Liu, Yingying Chen, and Marco Gruteser. 2010. Accuracy characterization of cell tower localization. In *UbiComp 2010: Ubiquitous Computing, 12th International Conference, UbiComp 2010, Copenhagen, Denmark, September 26-29, 2010, Proceedings.* 223–226.

[25] Zengbin Zhang et al. 2011. I Am the Antenna: Accurate Outdoor AP Location using Smartphones. In *Proc. ACM MobiCom.*